Vitesse de convergence et borne d'erreur pour l'algorithme LSTD(λ)

Manel Tagorti, Bruno Scherrer

Vitesse de convergence et borne d'erreur pour l'algorithme LSTD(λ)

Manel Tagorti et Bruno Scherrer

INRIA
615, rue du Jardin Botanique, 54600 Villers-lès-Nancy.
{manel.tagorti, bruno.scherrer}@inria.fr

Résumé : On considère l'algorithme LSTD(λ) (least-squares temporal-difference) avec traces d'éligibilité proposé par Boyan (2002). Cet algorithme renvoie, pour une politique fixée, une approximation linéaire de la fonction de valeur v pour les processus décisionnels de Markov admettant un grand espace d'états. On se restreint dans cet article au cas des chaînes de Markov β-mélangeantes. Sous cette hypothèse, on estime la vitesse de convergence de cet algorithme pour n'importe quelle valeur de $\lambda \in (0, 1)$. La borne d'erreur obtenue étend et améliore celle introduite par Lazaric *et al.* (2012) pour le cas $\lambda = 0$. L'analyse proposée permet de quantifier l'influence du paramètre λ, de l'espace linéaire de projection et du nombre d'échantillons utilisés.

Mots-clés : LSTD(λ), processus décisionnel de Markov, vitesse de convergence, bornes sur les performances.

1 Introduction

On considère dans cet article l'algorithme LSTD(λ) (least-squares temporal-difference) avec traces d'éligibilité proposé par Boyan (2002) dans le cadre de processus décionnels de Markov admettant un grand nombre d'états. C'est un algorithme couramment utilisé pour estimer la projection linéaire de la fonction de valeur étant donnée une politique fixée π. Il peut être particulièrement utile dans un schéma de type itérations sur les politiques pour approcher le contrôleur optimal (Bertsekas & Tsitsiklis, 1996; Szepesvári, 2010). Nedic & Bertsekas (2002) ont prouvé la convergence presque sûre de cet algorithme. Plus récemment, dans le cas $\lambda = 0$ et pour un nombre fini n d'échantillons, Lazaric *et al.* (2012) ont otbenu une borne d'erreur de l'ordre $\tilde{O}(\frac{1}{\sqrt{n}})$ [1] qui est valide avec forte probabilité. Une analyse similaire dans le cas $\lambda > 0$ n'a, à notre connaissance, pas encore été proposée dans la littérature ; c'est l'objet de cet article. Étudier le cas $\lambda \neq 0$ peut être intéréssant car ce paramètre intervient dans la qualité de la borne d'erreur entre la valeur asymptotique calculée par l'algorithme et la valeur effective v. En effet, comme le montre le théorème 2 (section 3) dû à Tsitsiklis & Roy (1997), en faisant varier λ de 0 à 1, on passe d'une façon continue d'une projection oblique (Scherrer, 2010) de v à sa projection orthogonale.

Cet article est organisé comme suit. La section 2 décrit le problème et l'algorithme considérés. La section 3 contient le résultat principal : pour tout $\lambda \in (0, 1)$, on montre que LSTD(λ) converge avec une vitesse de l'ordre $\tilde{O}(\frac{1}{\sqrt{n}})$. On en déduira une borne sur l'erreur globale (corollaire 1) qui nous renseigne sur le rôle du paramètre λ. La section 4 décrit les arguments de preuve de notre résultat. Finalement, la section 5 conclut et présente quelques perspectives.

2 Problématique et algorithme LSTD(λ)

On considère une chaîne de Markov $\mathcal{M}$ prenant ses valeurs dans un espace fini ou dénombrable [2]. On suppose $\mathcal{M}$ ergodique [3] ; Par conséquent elle admet une unique mesure invariante qu'on notera μ. Pour

1. La notation $f(n) = \tilde{O}(g(n))$ signifie $f(n) = O(g(n) \log^k g(n))$ pour $k \geq 0$.

2. On se restreint ici au cas fini/dénombrable car ceci facilite notre analyse. Une extension au cas continu est à notre avis possible moyennant quelques hypothèses techniques supplémentaires.

3. Dans le cas dénombrable, une chaîne de Markov est ergodique ssi elle est irréductible et apériodique, c'est-à-dire ssi : $\forall (x, y) \in \mathcal{X}^2,\ \exists n_0,\ \forall n \geq n_0,\ P^n(x, y) > 0$.

tout $K \in \mathbb{R}$, on note $\mathcal{B}(\mathcal{X}, K)$ l'ensemble des fonctions mesurables définies sur $\mathcal{X}$ et majorées par K. On considère la fonction récompense $r \in \mathcal{B}(\mathcal{X}, R_{\max})$ pour $R_{\max} \in \mathbb{R}$, qui renvoie la récompense reçue dans un état donné. La fonction de valeur v relative à la chaîne de Markov $\mathcal{M}$ est définie pour chaque état i comme l'espérance conditionnelle de la somme des récompenses cumulées le long d'une trajectoire infinie, sachant que l'état initial est i :

$$\forall i \in \mathcal{X},\ v(i) = \mathbb{E}\left[\sum_{j=0}^{\infty} \gamma^j r(X_j) \,\middle|\, X_0 = i\right],$$

où $\gamma \in (0,1)$ est le facteur d'actualisation. La fonction de valeur v est l'unique point fixe de l'opérateur de Bellman T :

$$\forall i \in \mathcal{X},\ Tv(i) = r(i) + \gamma \mathbb{E}\left[v(X_1) | X_0 = i\right].$$

Il est clair que $v \in \mathcal{B}(\mathcal{X}, V_{\max})$ avec $V_{\max} = \frac{R_{\max}}{1-\gamma}$. Quand l'espace des états $\mathcal{X}$ admet un grand nombre d'états, on peut approcher v en utilisant une *architecture d'approximation linéaire*. Pour cela, on va considérer la matrice des *features* Φ de dimension $|\mathcal{X}| \times d$, avec $d \ll |\mathcal{X}|$. Pour tout $x \in \mathcal{X}$, $\phi(x) = (\phi_1(x), ..., \phi_d(x))^T$ est *le vecteur feature* relatif à l'état x. On suppose que, pour tout $j \in \{1, ..., d\}$, la *fonction feature* $\phi_j : \mathcal{X} \mapsto \mathbb{R}$ appartient à $\mathcal{B}(\mathcal{X}, L)$, pour un L fini. On fera de plus l'hypothèse suivante [4].

Hypothèse 1. *Les fonctions features $(\phi_j)_{j \in \{1,...,d\}}$ sont linéairement indépendantes.*

Soit $\mathcal{S}$ le sous-espace vectoriel généré par $(\phi_j)_{1 \leq j \leq d}$. On considère la projection Π sur l'espace $\mathcal{S}$ suivant la norme quadratique $\|.\|_\mu$ définie pour tout f par

$$\|f\|_\mu = \sqrt{\sum_{x \in \mathcal{X}} |f(x)|^2 \mu(x)}.$$

La matrice de projection Π a la forme analytique suivante :

$$\Pi = \Phi(\Phi^T D_\mu \Phi)^{-1} \Phi^T D_\mu, \tag{1}$$

où D_μ est la matrice diagonale composée par les éléments $\mu(i)$. Le but de l'algorithme LSTD(λ) est de calculer la solution de l'équation $v = \Pi T^\lambda v$, où l'opérateur T^λ est défini comme la moyenne pondérée des puissances T^i de l'opérateur de Bellman T :

$$\forall \lambda \in (0,1),\ \forall v,\ T^\lambda v = (1-\lambda) \sum_{i=0}^{\infty} \lambda^i T^{i+1} v. \tag{2}$$

Dans le cas $\lambda = 0$, on a $T^\lambda = T$. On peut montrer que l'opérateur T^λ est contractant de module $\frac{(1-\lambda)\gamma}{1-\lambda\gamma} \leq \gamma$; en effet, comme T^i est affine et $\|P\|_\mu = 1$ (Tsitsiklis & Roy, 1997; Nedic & Bertsekas, 2002), pour tous vecteurs u et v sur $\mathcal{S}$, on a

$$\begin{aligned}
\|T^\lambda u - T^\lambda v\|_\mu &\leq (1-\lambda) \| \sum_{i=0}^{\infty} \lambda^i (T^{i+1} u - T^{i+1} v) \|_\mu \\
&= (1-\lambda) \| \sum_{i=0}^{\infty} \lambda^i (\gamma^{i+1} P^{i+1} u - \gamma^{i+1} P^{i+1} v) \|_\mu \\
&\leq (1-\lambda) \sum_{i=0}^{\infty} \lambda^i \gamma^{i+1} \|u - v\|_\mu \\
&= \frac{(1-\lambda)\gamma}{1-\lambda\gamma} \|u - v\|_\mu.
\end{aligned}$$

On sait que $\|\Pi\|_\mu = 1$ (Tsitsiklis & Roy, 1997), et on a par conséquent $\|\Pi T^\lambda\|_\mu < 1$. Par le théorème du point fixe de Banach, l'équation $v = \Pi T^\lambda v$ admet une et une seule solution, qu'on notera $v_{LSTD(\lambda)}$,

4. Cette hypothèse n'est pas centrale : en théorie, on peut enlever toutes les fonctions features qui rendent la famille (ϕ_i) non libre ; en pratique, on peut utiliser dans l'algorithme que nous allons décrire le pseudo-inverse au lieu de l'inverse.

étant donné que c'est la valeur vers laquelle l'algorithme LSTD(λ) converge asymptotiquement (Nedic & Bertsekas, 2002). Puisque $v_{LSTD(\lambda)}$ appartient au sous-espace $\mathcal{S}$, il existe $\theta \in \mathbb{R}^d$ tel que :

$$v_{LSTD(\lambda)} = \Phi\theta = \Pi T^\lambda \Phi\theta.$$

En remplaçant Π et T^λ par leurs expressions respectives (les équations 1 et 2), on peut montrer (Nedic & Bertsekas, 2002) que résoudre $v = \Pi T^\lambda v$ revient à résoudre l'équation $A\theta = b$, avec pour tout i,

$$A = \Phi^T D_\mu (I - \gamma P)(I - \lambda\gamma P)^{-1}\Phi = \mathbb{E}_{X_{-\infty}\sim\mu}\left[\sum_{k=-\infty}^{i} (\gamma\lambda)^{i-k}\phi(X_k)(\phi(X_i) - \gamma\phi(X_{i+1}))^T\right] \quad (3)$$

$$\text{et} \quad b = \Phi^T D_\mu (I - \gamma\lambda P)^{-1} r = \mathbb{E}_{X_{-\infty}\sim\mu}\left[\sum_{k=-\infty}^{i} (\gamma\lambda)^{i-k}\phi(X_k) r(X_i)\right], \quad (4)$$

où on note u^T la transposée de u. Pour tout x, $\phi(x)$ est de dimension d, donc la matrice A est une matrice de taille $d \times d$ et b un vecteur de taille d. Sous l'hypothèse 1, on peut montrer que la matrice A est inversible (Nedic & Bertsekas, 2002) et donc $v_{LSTD(\lambda)} = \Phi A^{-1} b$ est bien défini.

On va maintenant décrire plus précisément l'algorithme LSTD(λ). Étant donnée une trajectoire $X_1, ..., X_n$ générée par la chaîne de Markov $\mathcal{M}$, les expressions respectives de A et b sous forme d'espérance (équations (3)-(4)) suggèrent de faire les estimations suivantes :

$$\hat{A} = \frac{1}{n-1}\sum_{i=1}^{n-1} z_i(\phi(X_i) - \gamma\phi(X_{i+1}))^T$$

$$\text{et} \quad \hat{b} = \frac{1}{n-1}\sum_{i=1}^{n-1} z_i r(X_i),$$

$$\text{où} \quad z_i = \sum_{k=1}^{i} (\lambda\gamma)^{i-k}\phi(X_k) \quad (5)$$

est appelée la *trace d'éligibilité*. L'algorithme LSTD(λ) renvoie alors l'estimation (pour un nombre d'échantillons fini) $\hat{v}_{LSTD(\lambda)} = \Phi\hat{\theta}$ de $v_{LSTD(\lambda)}$ avec[5] $\hat{\theta} = \hat{A}^{-1}\hat{b}$. Nedic & Bertsekas (2002) ont montré, en utilisant une variante de la loi des grands nombres, la convergence presque sûre de $\hat{A}$ vers A et $\hat{b}$ vers b. Ceci implique la convergence presque sûre de $\hat{v}_{LSTD(\lambda)}$ vers $v_{LSTD(\lambda)}$. Le but de cet article est d'approfondir cette analyse : on va estimer la vitesse de convergence de $\hat{v}_{LSTD(\lambda)}$ vers $v_{LSTD(\lambda)}$, et borner l'erreur d'approximation $\|\hat{v}_{LSTD(\lambda)} - v\|_\mu$ de l'algorithme.

3 Résultat principal

Cette section contient notre résultat principal. L'hypothèse clé dans notre analyse consiste à supposer la chaîne de Markov β-mélangeante[6].

Hypothèse 2. *Le processus $(X_n)_{n\geq 1}$ est β-mélangeant, c'est-à-dire que son $i^{ème}$ coefficient*

$$\beta_i = \sup_{t\geq 1} \mathbb{E}\left[\sup_{B\in\sigma(X_{t+i}^{\infty})} \left|P(B|\sigma(X_1^t)) - P(B)\right|\right]$$

tend vers 0 quand i tend vers l'infini, avec $X_l^j = \{X_l, ..., X_j\}$ pour $j \geq l$ et $\sigma(X_l^j)$ la sigma-algèbre générée par X_l^j. De plus, le processus $(X_n)_{n\geq 1}$ est exponentiellement β-mélangeant de paramètres $\overline{\beta} > 0$, $b > 0$, et $\kappa > 0$ dans le sens suivant : $\beta_i \leq \overline{\beta}e^{-bi^\kappa}$.

Intuitivement le coefficient β_i mesure le degré de dépendance entre les échantillons de la séquence séparés par i pas de temps (plus le coefficient est petit, plus les échantillons sont indépendants). On va maintenant énoncer le théorème principal qui fournit la vitesse de convergence de l'algorithme LSTD(λ).

5. On verra dans le théorème 1 que $\hat{A}$ est inversible avec forte probabilité à partir d'un certain nombre d'échantillons.

6. Une chaîne de Markov stationnaire et ergodique est *toujours* β-mélangeante.

Théorème 1. *On fait les hypothèses 1 et 2 et on suppose que $X_1 \sim \mu$. Pour tout $n \geq 1$ et $\delta \in (0,1)$, on définit les fonctions :*

$$I(n,\delta) = 32\Lambda(n,\delta)\max\left\{\frac{\Lambda(n,\delta)}{b}, 1\right\}^{\frac{1}{\kappa}}$$
$$\text{où } \Lambda(n,\delta) = 2\left(\log\left(\frac{8n^2}{\delta}\right) + \log(\max\{4e^2, n\overline{\beta}\})\right).$$

Soit $n_0(\delta)$ le plus petit entier tel que pour tout $n \geq n_0(\delta)$,

$$\frac{2dL^2}{(1-\gamma)\nu}\left[\frac{2}{\sqrt{n-1}}\sqrt{\left(\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil+1\right)I(n-1,\delta)} + \frac{1}{(n-1)(1-\lambda\gamma)} + \frac{2}{n-1}\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\right] < 1 \tag{6}$$

où ν est la plus petite valeur propre de la matrice de Gram $\Phi^T D_\mu \Phi$. Alors, pour tout δ, avec une probabilité supérieure ou égale à $1-\delta$, on a pour tout $n \geq n_0(\delta)$, $\hat{A}$ inversible et :

$$\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_\mu \leq \frac{4V_{max}dL^2}{\sqrt{n-1}(1-\gamma)\nu}\sqrt{\left(1+\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\right)I(n-1,\delta)} + h(n,\delta)$$

avec $h(n,\delta) = \tilde{O}(\frac{1}{n})$.

La constante ν est strictement positive sous l'hypothèse 1. Pour tout δ, il est clair que l'entier $n_0(\delta)$ existe et est fini puisque le terme à gauche tend vers 0 quand n tend vers l'infini. Comme la fonction $\left(1+\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\right)I(n-1,\frac{\delta}{n^2})$ vérifie $\left(1+\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\right)I(n-1,\frac{\delta}{n^2}) = \tilde{O}(1)$, on peut déduire que LSTD(λ) estime $v_{LSTD(\lambda)}$ avec une vitesse de l'ordre $\tilde{O}\left(\frac{1}{\sqrt{n}}\right)$. Comme la fonction $\lambda \mapsto \frac{1}{\log\left(\frac{1}{\lambda\gamma}\right)}$ est croissante, notre borne sur la vitesse de convergence a tendance à se détériorer lorsque λ augmente. D'autre part, la garantie de qualité de $v_{LSTD(\lambda)}$ s'améliore lorsqu'on augmente le paramètre λ, comme le montre le résultat suivant de la littérature.

Théorème 2 (Tsitsiklis & Roy (1997))**.** *L'erreur d'approximation vérifie*[7] *:*

$$\|v - v_{LSTD(\lambda)}\|_\mu \leq \frac{1-\lambda\gamma}{1-\gamma}\|v - \Pi v\|_\mu.$$

En particulier lorsque $\lambda = 1$, on a $\frac{1-\lambda\gamma}{1-\gamma} = 1$ et l'on en déduit que LSTD(1) estime la projection orthogonale Πv de v. En utilisant l'inégalité triangulaire, on déduit des théorèmes 1 et 2 une borne sur l'erreur globale.

Corollaire 1. *Sous les mêmes hypothèses et les mêmes les notations du théorème 1, pour tout δ, avec une probabilité $1-\delta$, pour tout $n \geq n_0(\delta)$, l'erreur globale de LSTD(λ) vérifie :*

$$\|v - \hat{v}_{LSTD(\lambda)}\|_\mu \leq \frac{1-\lambda\gamma}{1-\gamma}\|v - \Pi v\|_\mu + \frac{4V_{max}dL^2}{\nu\sqrt{n-1}(1-\gamma)}\left(\left(1+\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\right)I(n-1,\delta)\right)^{\frac{1}{2}} + h(n,\delta).$$

Remarque 1. *Le résultat énoncé dans le corollaire 1 est légèrement plus fort que celui de Lazaric* et al. *(2012) dans le cas $\lambda = 0$: Pour une propriété $P(n)$, notre résultat est de la forme "$\forall\delta, \exists n_0(\delta)$ tq. $\forall n \geq n_0(\delta)$, $P(n)$ est vraie avec probabilité $1-\delta$" alors que le leur est de la forme "$\forall n$, $\forall\delta$, $P(n)$ est vraie*

7. Cette borne peut être améliorée, comme l'a suggéré V. Papavassilou (Tsitsiklis & Roy, 1997) ; en utilisant le théorème de Pythagore, on obtient

$$\|v - v_{LSTD(\lambda)}\|_\mu \leq \frac{1-\lambda\gamma}{\sqrt{(1-\gamma)(1+\gamma-2\lambda\gamma)}}\|v - \Pi v\|_\mu.$$

Par souci de simplicité, nous garderons la forme du théoreme 2.

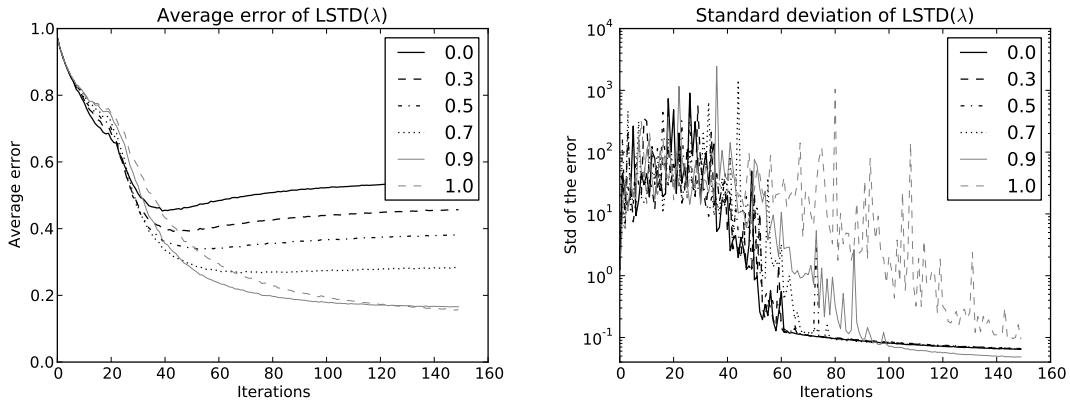

FIGURE 1 – **Courbes d'apprentissage pour différentes valeurs de** λ. On génère 1000 MDPs Garnet aléatoires (Archibald *et al.*, 1995) avec 100 états, des récompenses uniformes et aléatoires avec $\gamma = 0.99$. On a aussi généré 1000 features d'espace aléatoires de dimension 20 (en prenant des matrices aléatoires avec des entrées uniformes et aléatoires). Pour toutes ces valeurs de $\lambda \in \{0.0, 0.3, 0.5, 0.7, 0.9, 1.0\}$, on montre (à gauche) la moyenne de l'erreur *réelle* et (à droite) la déviation standard en fonction du nombre d'échantillons. Empiriquement, la meilleure valeur de λ a l'air d'être une fontion monotone du nombre d'échantillons n, qui tend vers 1 asymptotiquement. Ceci concorde avec le résultat énoncé dans le Corollaire 1.

avec probabilité $1 - \delta$*". De plus, sous les mêmes hypothèses, l'erreur globale obtenue par Lazaric* et al. *(2012), est bornée par*

$$\|\tilde{v}_{LSTD(0)} - v\|_\mu \leq \frac{4\sqrt{2}}{1-\gamma}\|v - \Pi v\|_\mu + \tilde{O}\left(\frac{1}{\sqrt{n}}\right),$$

où $\tilde{v}_{LSTD(0)}$ *est la solution tronquée (avec* V_{max}*) de l'algorithme pathwise LSTD*[8]*, alors qu'avec notre analyse nous obtenons*

$$\|\hat{v}_{LSTD(0)} - v\|_\mu \leq \frac{1}{1-\gamma}\|v - \Pi v\|_\mu + \tilde{O}\left(\frac{1}{\sqrt{n}}\right).$$

Le terme correspondant à l'erreur d'approximation est plus fin d'un facteur $4\sqrt{2}$ *avec notre analyse. Aussi, contrairement à ce que l'on présente ici, l'analyse proposée par (Lazaric* et al.*, 2012) ne nous renseigne pas sur la vitesse de convergence de LSTD(*0*) (ils ne proposent pas une borne sur l'erreur* $\|v_{LSTD(0)} - \hat{v}_{LSTD(0)}\|_\mu$*). Leur analyse est basée sur un modèle de régression Markovien qui consiste à borner* directement *l'erreur globale selon la norme* μ*. Notre argumentation, qui consiste à borner l'erreur d'approximation puis l'erreur d'estimation, permet d'avoir un résultat plus fin.*

Comme il a déjà été mentionné précédemment, la valeur $\lambda = 1$ minimise la borne sur l'erreur d'approximation $\|v - v_{LSTD(\lambda)}\|_\mu$ (le premier terme à droite dans le corollaire 1) alors que la valeur $\lambda = 0$ minimise la borne sur l'erreur d'estimation $\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_\mu$ (le second terme). Il existe pour tout n et pour tout δ une valeur optimale λ^* du paramètre qui minimise l'erreur globale, créant ainsi un compromis entre ces deux erreurs. La figure 1 illustre la relation qui existe entre λ et n. Le choix optimal λ^* depend aussi bien des paramètres du processus β-mélangeant (b, κ et $\overline{\beta}$) que de la qualité de l'espace de projection $\|v - \Pi v\|_\mu$, qui sont généralement des quantités inconnues en pratique. Il est clair cependant que lorsque n tend vers l'infini, λ^* tend vers 1.

La section suivante contient une preuve détaillée du théorème 1.

4 Preuve du théorème 1

Dans cette section, on développe les arguments justifiant les résultats de la section précédente. La preuve est organisée en deux parties. Dans la première partie, on prouve une inégalité de concentration pour des

8. Voir (Lazaric *et al.*, 2012) pour plus de details.

processus vectoriels avec traces d'éligibilité infiniment longues. Ensuite dans la seconde partie, on prouve le théorème 1 : on applique l'inégalité à l'erreur induite en estimant A et b, puis on relie ces erreurs à celle sur $v_{LSTD(\lambda)}$.

4.1 Inégalité de concentration pour les estimations avec des traces d'éligiblité infiniment longues

L'une des premières difficultés de l'algorithme LSTD(λ) est que les variables $A_i = z_i(\phi(X_i) - \gamma\phi(X_{i+1}))^T$ (respectivement $b_i = z_i r(X_i)$) ne sont pas indépendantes. On ne peut plus dès lors appliquer les résultats standards de concentration (comme celui décrit dans le lemme 6 dans l'annexe A) afin d'estimer la vitesse de convergence des estimations vers leur limites. Puisque les variables A et b admettent la même structure, on notera G la matrice qui a cette forme générale

$$\hat{G} = \frac{1}{n-1}\sum_{i=1}^{n-1} G_i \tag{7}$$

$$\text{avec} \quad G_i = z_i(\tau(X_i, X_{i+1}))^T \tag{8}$$

où z_i, défini dans l'équation (5), satisfait $z_i = \sum_{k=1}^{i}(\lambda\gamma)^{i-k}\phi(X_k)$, et $\tau : \mathcal{X}^2 \mapsto \mathbb{R}^k$ est tel que, pour tout $1 \leq i \leq k$, τ_i appartient à $\mathcal{B}(\mathcal{X}^2, L')$ pour une constante L' finie[9]. Les variables G_i sont calculées à partir d'une même trajectoire, et sont par conséquent significativement dépendantes. Néanmoins, grâce à l'hypothèse 2, nous allons être en mesure d'utiliser la technique de blocs de Yu (1994) qui nous ramène au cas indépendant. Ce passage du cas β-mélangeant au cas indépendant nécessite l'hypothèse de stationnarité (lemme 5). Malheureusement, les variables G_i ne définissent pas un processus stationnaire puisqu'elles sont une fonction $\sigma(\mathcal{X}^{i+1})$-mesurable du vecteur *non-stationnaire* $(X_1, \ldots, X_{i+1})$. Afin de résoudre ce problème, on va approcher G_i par sa version tronquée stationnaire G_i^m. Ceci est possible si on approche la trace z_i par sa version m-tronquée :

$$z_i^m = \sum_{k=\max(i-m+1,1)}^{i} (\lambda\gamma)^{i-k}\phi(X_k).$$

Puisque la fonction ϕ est bornée par une certaine constante L et que l'influence des anciens événements est bornée par une puissance de $\lambda\gamma < 1$, on peut montrer que $\|z_i - z_i^m\|_\infty \leq \frac{L}{1-\lambda\gamma}(\lambda\gamma)^m$. Pour m vérifiant $m > \frac{\log(n-1)}{\log\frac{1}{\lambda\gamma}}$, on obtient $\|z_i - z_i^m\|_2 = O\left(\frac{1}{n}\right)$. Il paraît alors raisonnable d'approcher G par $\hat{G}^m$ satisfaisant

$$\hat{G}^m = \frac{1}{n-1}\sum_{i=1}^{n-1} G_i^m, \tag{9}$$

$$\text{avec} \quad G_i^m = z_i^m(\tau(X_i, X_{i+1}))^T. \tag{10}$$

Pour tout $i \geq m$, G_m^i est une fonction $\sigma(\mathcal{X}^{m+1})$-mesurable du vecteur stationnaire $Z_i = (X_{i-m+1}, X_{i-m+2}, \ldots, X_{i+1})$. On peut alors appliquer la technique de Yu (1994) à G_i^m, mais pour cela il faut d'abord vérifier que les variables G_i^m définissent bien un processus β-mélangeant. On sait d'après Yu (1994) que toute fonction d'un processus β-mélangeant est un processus β-mélangeant de coefficient $\beta^f \leq \beta$, donc il nous suffit de prouver que le processus Z_i est β-mélangeant. Pour cela on va essayer de relier les coefficients β_Z du processus Z_i à ceux du processus X_n supposé β-mélangeant d'après l'hypothèse 2. C'est ce que l'on fait dans le lemme suivant.

Lemme 1. *Soit $(X_n)_{n\geq 1}$ un processus β-mélangeant alors $(Z_n)_{n\geq 1} = (X_{n-m+1}, X_{n-m+2}, \ldots, X_{n+1})_{n\geq 1}$ est un processus β-mélangeant dont le i-ème coefficient β-mélangeant β_i^Z vérifie $\beta_i^Z \leq \beta_{i-m}^X$.*

Démonstration. Soit $\Gamma = \sigma(Z_1, ..., Z_t)$, par définition on a

$$\Gamma = \sigma(Z_j^{-1}(B) : j \in \{1, ..., t\}, B \in \sigma(\mathcal{X}^{m+1})).$$

9. On note $\mathcal{X}^i = \underbrace{X \times \mathcal{X} ... \times \mathcal{X}}_{\text{i fois}}$ pour $i \geq 1$.

Pour tout $j \in \{1, ..., t\}$ on a

$$Z_j^{-1}(B) = \{\omega \in \Omega, Z_j(\omega) \in B\}.$$

Pour $B = B_0 \times ... \times B_m$, on peut observer que

$$Z_j^{-1}(B) = \{\omega \in \Omega, X_j(\omega) \in B_0, ..., X_{j+m}(\omega) \in B_m\}.$$

D'où, on peut s'apercevoir que

$$\Gamma = \sigma(X_j^{-1}(B) : j \in \{1, ..., t+m\}, B \in \sigma(\mathcal{X})) = \sigma(X_1, ..., X_{t+m}).$$

De même on peut montrer que $\sigma(Z_{t+i}^{\infty}) = \sigma(X_{t+i}^{\infty})$. Soit β_i^X le $i^{\text{ème}}$ coefficient β-mélangeant du processus $(X_n)_{n\geq 1}$. On a

$$\beta_i^X = \sup_{t\geq 1} \mathbb{E}\left[\sup_{B\in\sigma(X_{t+i}^{\infty})} |P(B|\sigma(X_1, ..., X_t)) - P(B)|\right].$$

D'une façon similaire, on a

$$\beta_i^Z = \sup_{t\geq 1} \mathbb{E}\left[\sup_{B\in\sigma(Z_{t+i}^{\infty})} |P(B|\sigma(Z_1, ..., Z_t)) - P(B)|\right].$$

En appliquant les identités qu'on a montrées juste avant on peut voir que

$$\beta_i^Z = \sup_{t\geq 1} \mathbb{E}\left[\sup_{B\in\sigma(X_{t+i}^{\infty})} |P(B|\sigma(X_1, ..., X_{t+m})) - P(B)|\right].$$

Notant $t' = t + m$, on a pour $i > m$

$$\begin{aligned}\beta_i^Z &= \sup_{t'\geq m+1} \mathbb{E}\left[\sup_{B\in\sigma(X_{t'+i-m}^{\infty})} |P(B|\sigma(X_1, ..., X_{t'})) - P(B)|\right] \\ &\leq \beta_{i-m}^X.\end{aligned}$$

□

Soit $\|.\|_F$ la norme de Frobenius : pour $M \in \mathbb{R}^{d\times k}$, $\|M\|_F^2 = \sum_{l=1}^{d}\sum_{j=1}^{k}(M_{l,j})^2$. On peut maintenant énoncer l'inégalité de concentration pour le processus β-mélangeant $\hat{G}$ avec des traces infiniment longues.

Lemme 2. *Soit la matrice G_i de taille $d \times k$ définie par*

$$G_i = \sum_{k=1}^{i} (\lambda\gamma)^{i-k}\phi(X_k)(\tau(X_i, X_{i+1}))^T. \tag{11}$$

Rappelons que $\phi = (\phi_1, ..., \phi_d)$ est tel que pour tout j, $\phi_j \in \mathcal{B}(X, L)$ et que $\tau \in \mathcal{B}(\mathcal{X}^2, L')$. Sous les hypothèses et notations du théorème 1, on a pour tout $\delta \in (0, 1)$, avec probabilité $1 - \delta$,

$$\left\|\frac{1}{n-1}\sum_{i=1}^{n-1} G_i - \frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{E}[G_i]\right\|_2 \leq \frac{B_2}{\sqrt{n-1}}\sqrt{(m+1)J(n-1,\delta)} + \epsilon(n),$$

où

$$\begin{aligned}J(n,\delta) &= 32\Gamma(n,\delta)\max\left\{\frac{\Gamma(n,\delta)}{b}, 1\right\}^{\frac{1}{\kappa}}, \\ \Gamma(n,\delta) &= \log\left(\frac{2}{\delta}\right) + \log(\max\{4e^2, n\overline{\beta}\}), \\ \epsilon(n) &= 2\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\frac{\sqrt{d\times k}LL'}{(n-1)(1-\lambda\gamma)}.\end{aligned}$$

Par rapport aux quantités I et Λ introduites dans l'énoncé du Théorème 1, les quantités introduites ici sont telles que $J(n,\delta) = I(n, 4n^2\delta)$ et $\Gamma(n,\delta) = \Lambda(n, 4n^2\delta)$.

Démonstration. La preuve consiste i) à montrer que l'erreur induite en considérant $\hat{G}^m$, la version avec la trace tronquée, au lieu de $\hat{G}$ est bornée par $\epsilon(n)$, et ii) à appliquer la technique de blocs de Yu (1994) similaire à ce que Lazaric *et al.* (2012) ont employée pour LSTD(0). On peut trouver une preuve détaillée de ce lemme dans l'annexe A. □

En utilisant une preuve fortement similaire, on peut déduire l'inégalité de concentration générale suivante pour les processus β-mélangeants.

Lemme 3. *Soient $Y = (Y_1, \ldots, Y_n)$ des variables aléatoires prenant leurs valeurs dans l'espace $\mathbb{R}^d$, générées par un processus exponentiellement β-mélangeant stationnaire avec paramètres $\overline{\beta}$, b et κ, tel que pour tout i, $\|Y_i - \mathbb{E}[Y_i]\|_2 \leq B_2$ presque sûrement. Alors pour tout $\delta > 0$,*

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i]\right\|_2 \leq \frac{B_2}{\sqrt{n}}\sqrt{J(n,\delta)}\right\} > 1-\delta.$$

où $J(n,\delta)$ est défini dans le Lemme 2.

Remarque 2. *Si les variables Y_i étaient indépendantes, on aurait $\beta_i = 0$ pour tout i, c'est-à-dire que, en prenant $\overline{\beta} = 0$ et $b = \infty$, $J(n,\delta)$ est dans ce cas égale à $32\log\frac{8e^2}{\delta} = O(1)$, et on retrouve bien le résultat standard décrit dans le lemme 6 (dans l'annexe A). Le prix à payer pour le fait d'avoir des échantillons β-mélangeants au lieu d'échantillons indépendants est le terme $J(n,\delta) = \tilde{O}(1)$; en d'autres termes, ce prix est raisonnable.*

4.2 Preuve du Théorème 1

Une fois qu'on a introduit le résultat de concentration, on peut prouver le Théorème 1. Une étape importante avant cela consiste à dériver le lemme suivant.

Lemme 4. *Soient $\epsilon_A = \hat{A} - A$ et $\epsilon_b = \hat{b} - b$. Soit ν la plus petite valeur propre de la matrice $\Phi^T D_\mu \Phi$. Pour tout $\lambda \in (0,1)$, l'estimation $\hat{v}_{LSTD(\lambda)}$ vérifie* [10] *:*

$$\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_\mu \leq \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}}\|(I+\epsilon_A A^{-1})^{-1}\|_2\|\epsilon_A\theta - \epsilon_b\|_2.$$

De plus, si les constantes ϵ et C sont telles que $\|\epsilon_A\|_2 \leq \epsilon < C \leq \frac{1}{\|A^{-1}\|}$, alors $\hat{A}$ est inversible et

$$\|(I+\epsilon_A A^{-1})^{-1}\|_2 \leq \frac{1}{1-\frac{\epsilon}{C}}.$$

Démonstration. À partir des définitions de $v_{LSTD(\lambda)}$ et $\hat{v}_{LSTD(\lambda)}$, on a

$$\begin{aligned}\hat{v}_{LSTD(\lambda)} - v_{LSTD(\lambda)} &= \Phi\hat{\theta} - \Phi\theta \\ &= \Phi A^{-1}(A\hat{\theta} - b). \end{aligned} \tag{12}$$

D'une part, on peut observer en partant de l'expression de A dans l'équation (3) et en écrivant $M = (1-\lambda)\gamma P(I-\lambda\gamma P)^{-1}$ et $M_\mu = \Phi^T D_\mu \Phi$ que

$$\begin{aligned}\Phi A^{-1} &= \Phi\left[\Phi^T D_\mu (I-\gamma P)(I-\lambda\gamma P)^{-1}\Phi\right]^{-1} \\ &= \Phi\left[\Phi^T D_\mu (I-\lambda\gamma P-(1-\lambda)\gamma P)(I-\lambda\gamma P)^{-1}\Phi\right]^{-1} \\ &= \Phi(M_\mu - \Phi^T D_\mu M\Phi)^{-1}.\end{aligned}$$

10. Quand $\hat{A}$ n'est pas inversible, on prend $\hat{v}_{LSTD(\lambda)} = \infty$ et l'inégalité est toujours vraie puisque, comme on va le voir, l'inversibilité de $\hat{A}$ est équivalente à celle de $(I+\epsilon_A A^{-1})$.

Les matrices A et M_μ sont inversibles donc la matrice $I - M_\mu^{-1}\Phi^T D_\mu M\Phi$ l'est également et on a

$$\Phi A^{-1} = \Phi(I - M_\mu^{-1}\Phi^T D_\mu M\Phi)^{-1}M_\mu^{-1}.$$

On sait d'après Tsitsiklis & Roy (1997) que $\|\Pi\|_\mu = 1$—on rappelle que Π est la projection définie à l'équation (1)—et $\|P\|_\mu = 1$. Il s'en suit que $\|\Pi M\|_\mu = \frac{(1-\lambda)\gamma}{1-\lambda\gamma} < 1$ et que la matrice $(I - \Pi M)$ est inversible. En utilisant l'identité $X(I - YX)^{-1} = (I - XY)^{-1}X$ avec $X = \Phi$ et $Y = M_\mu^{-1}\Phi^T D_\mu M$, on obtient

$$\Phi A^{-1} = (I - \Pi M)^{-1}\Phi M_\mu^{-1}. \tag{13}$$

D'autre part, en utilisant les identités $A\theta = b$ et $\hat{A}\hat{\theta} = \hat{b}$ on peut voir que pour $\epsilon_A = \hat{A} - A$ on a :

$$\begin{aligned}
A\hat{\theta} - b &= A\hat{\theta} - b - (\hat{A}\hat{\theta} - \hat{b}) \\
&= \hat{b} - b - \epsilon_A\theta + \epsilon_A\theta - \epsilon_A\hat{\theta} \\
&= \hat{b} - b - (\hat{A} - A)\theta + \epsilon_A(\theta - \hat{\theta}) \\
&= \hat{b} - \hat{A}\theta - (b - A\theta) + \epsilon_A A^{-1}(A\theta - A\hat{\theta}) \\
&= \hat{b} - \hat{A}\theta + \epsilon_A A^{-1}(b - A\hat{\theta}).
\end{aligned}$$

D'où

$$A\hat{\theta} - b = \hat{b} - \hat{A}\theta - \epsilon_A A^{-1}(A\hat{\theta} - b).$$

Ceci implique que

$$\begin{aligned}
A\hat{\theta} - b &= (I + \epsilon_A A^{-1})^{-1}(\hat{b} - \hat{A}\theta) \\
&= (I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta).
\end{aligned} \tag{14}$$

où la deuxième égalité vient de l'identité $A\theta = b$. En utilisant les équations (13) et (14), l'équation (12) peut être réécrite de la manière suivante :

$$\hat{v}_{LSTD(\lambda)} - v_{LSTD(\lambda)} = (I - \Pi M)^{-1}\Phi M_\mu^{-1}(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta). \tag{15}$$

On va maintenant essayer de borner le terme $\|\Phi M_\mu^{-1}(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta)\|_\mu$. Pour tout x, on a

$$\|\Phi M_\mu^{-1}x\|_\mu = \sqrt{x^T M_\mu^{-1}\Phi^T D_\mu \Phi M_\mu^{-1}x} = \sqrt{x^T M_\mu^{-1}x} \leq \frac{1}{\sqrt{\nu}}\|x\|_2, \tag{16}$$

où ν est plus petite valeur propre réelle de la matrice de Gram M_μ. En prenant la norme $\|.\|_\mu$ dans l'équation (15) et en utilisant l'inégalité précédente, on obtient :

$$\begin{aligned}
\|\hat{v}_{LSTD(\lambda)} - v_{LSTD(\lambda)}\|_\mu &\leq \|(I - \Pi M)^{-1}\|_\mu \|\Phi M_\mu^{-1}(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta)\|_\mu \\
&\leq \|(I - \Pi M)^{-1}\|_\mu \frac{1}{\sqrt{\nu}}\|(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta)\|_2 \\
&\leq \|(I - \Pi M)^{-1}\|_\mu \frac{1}{\sqrt{\nu}}\|(I + \epsilon_A A^{-1})^{-1}\|_2\|(\epsilon_b - \epsilon_A\theta)\|_2.
\end{aligned}$$

La première partie du lemme est obtenue en utilisant à nouveau l'égalité $\|\Pi M\|_\mu = \frac{(1-\lambda)\gamma}{1-\lambda\gamma} < 1$, qui implique que

$$\|(I - \Pi M)^{-1}\|_\mu = \left\|\sum_{i=0}^{\infty}(\Pi M)^i\right\|_\mu \leq \sum_{i=0}^{\infty}\|\Pi M\|_\mu^i \leq \frac{1}{1 - \frac{(1-\lambda)\gamma}{1-\lambda\gamma}} = \frac{1-\lambda\gamma}{1-\gamma}. \tag{17}$$

On va maintenant prouver la seconde partie du lemme. La matrice $\hat{A}$ est inversible si et seulement si la matrice $\hat{A}A^{-1} = (A + \epsilon_A)A^{-1} = I + \epsilon_A A^{-1}$ l'est également. On note $\rho(\epsilon_A A^{-1})$ le rayon spectral de la matrice $\epsilon_A A^{-1}$. Une condition suffisante pour l'inversibilité de $\hat{A}A^{-1}$ est d'imposer que $\rho(\epsilon_A A^{-1}) < 1$. On

a pour toute matrice M réelle carrée $\rho(M) \leq \|M\|_2$. Ainsi, pour ϵ et C tels que $\|\epsilon_A\|_2 \leq \epsilon < C < \frac{1}{\|A^{-1}\|_2}$, on déduit

$$\rho(\epsilon_A A^{-1}) \leq \|\epsilon_A A^{-1}\|_2 \leq \|\epsilon_A\|_2 \|A^{-1}\|_2 \leq \frac{\epsilon}{C} < 1.$$

La matrice $\hat{A}$ est alors inversible et on a :

$$\|(I + \epsilon_A A^{-1})^{-1}\|_2 = \left\| \sum_{i=0}^{\infty} (\epsilon_A A^{-1})^i \right\|_2 \leq \sum_{i=0}^{\infty} \left(\frac{\epsilon}{C}\right)^i = \frac{1}{1 - \frac{\epsilon}{C}}.$$

Ceci conclut la preuve du lemme 4. □

Il suffit d'après le lemme 4, pour conclure la preuve du théorème 1, de contrôler les termes $\|\epsilon_A\|_2$ et $\|\epsilon_A \theta - \epsilon_b\|_2$. C'est ce que l'on fait maintenant.

Contrôle du terme $\|\epsilon_A\|_2$.

En utilisant l'inégalité triangulaire, on peut voir que

$$\|\epsilon_A\|_2 \leq \|\mathbb{E}[\epsilon_A]\|_2 + \|\epsilon_A - \mathbb{E}[\epsilon_A]\|_2. \tag{18}$$

Soit $\hat{A}_{n,k} = \phi(X_k)(\phi(X_n) - \gamma\phi(X_{n+1}))^T$. Pour tout n et k, on a $\|\hat{A}_{n,k}\|_2 \leq 2dL^2$. On peut borner le premier terme à droite de l'inégalité dans l'équation (18) comme suit :

$$\begin{aligned}
\|\mathbb{E}[\epsilon_A]\|_2 &= \left\| A - \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1}\sum_{k=1}^{i}(\lambda\gamma)^{i-k}\hat{A}_{i,k}\right]\right\|_2 \\
&= \left\| \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1}\left(\sum_{k=-\infty}^{i}(\lambda\gamma)^{i-k}\hat{A}_{i,k} - \sum_{k=1}^{i}(\lambda\gamma)^{i-k}\hat{A}_{i,k}\right)\right]\right\|_2 \\
&= \left\| \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1}(\lambda\gamma)^{i}\sum_{k=-\infty}^{0}(\lambda\gamma)^{-k}\hat{A}_{i,k}\right]\right\|_2 \\
&\leq \frac{1}{n-1}\sum_{i=1}^{n-1}(\lambda\gamma)^i \frac{2dL^2}{1-\lambda\gamma} \\
&\leq \frac{1}{n-1}\frac{2dL^2}{(1-\lambda\gamma)^2} = \epsilon_0(n).
\end{aligned}$$

Soit δ_n un paramètre dans $(0, 1)$, qui dépend de n et qu'on fixera par la suite. A partir de l'équation (18) et de la borne déduite, on obtient :

$$\begin{aligned}
\mathbb{P}(\|\epsilon_A\|_2 \geq \epsilon_1(n, \delta_n)) &\leq \mathbb{P}(\|\epsilon_A - \mathbb{E}[\epsilon_A]\|_2 \geq \epsilon_1(n, \delta_n) - \epsilon_0(n)) \\
&\leq \delta_n
\end{aligned}$$

pour $\epsilon_1(n, \delta_n)$ vérifiant—cf. lemme 2—$\epsilon_1(n, \delta_n) - \epsilon_0(n) = \frac{4dL^2}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{(m+1)J(n-1, \delta_n)} + \epsilon(n)$ avec $\epsilon(n) = \frac{4mdL^2}{(n-1)(1-\lambda\gamma)}$, c'est-à-dire

$$\epsilon_1(n, \delta_n) = \frac{4dL^2}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{(m+1)J(n-1, \delta_n)} + \epsilon(n) + \epsilon_0(n). \tag{19}$$

Contrôle du terme $\|\epsilon_A\theta - \epsilon_b\|_2$.

En utilisant le fait que $A\theta = b$, les définitions de $\hat{A}$ et $\hat{b}$, et le fait que $\phi(x)^T\theta = [\phi\theta](x)$, on a

$$\begin{aligned}\epsilon_A\theta - \epsilon_b &= \hat{A}\theta - \hat{b} \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1} z_i(\phi(X_i) - \gamma\phi(X_{i+1})^T)\theta - \frac{1}{n-1}\sum_{i=1}^{n-1} z_i r(X_i) \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1} z_i([\phi\theta](X_i) - \gamma[\phi\theta](X_{i+1})^T - r(X_i)) \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1} z_i\Delta_i\end{aligned}$$

où, comme $v_{LSTD(\lambda)} = \Phi\theta$, Δ_i est un nombre égal à

$$\Delta_i = v_{LSTD(\lambda)}(X_i) - \gamma v_{LSTD(\lambda)}(X_{i+1}) - r(X_i).$$

On peut ainsi contrôler $\|\epsilon_A\theta - \epsilon_b\|_2$ en suivant les mêmes étapes de preuve que précédemment. En effet, on a

$$\begin{aligned}&\|\epsilon_A\theta - \epsilon_b\|_2 \leq \|\epsilon_A\theta - \epsilon_b - \mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 + \|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \qquad (20)\\ \text{et}\quad &\|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \leq \|\mathbb{E}[\epsilon_A]\|_2\|\theta\|_2 + \|\mathbb{E}[\epsilon_b]\|_2.\end{aligned}$$

On a $\|\mathbb{E}[\epsilon_A]\|_2 \leq \epsilon_0(n) = \frac{1}{n-1}\frac{2dL^2}{(1-\lambda\gamma)^2}$. On peut montrer similairement que $\|\mathbb{E}[\epsilon_b]\|_2 \leq \frac{1}{n-1}\frac{\sqrt{d}LR_{\max}}{(1-\lambda\gamma)^2}$. On peut donc conclure que

$$\|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \leq \frac{1}{n-1}\frac{2dL^2}{(1-\lambda\gamma)^2}\|\theta\|_2 + \frac{1}{n-1}\frac{\sqrt{d}LR_{\max}}{(1-\lambda\gamma)^2} = \epsilon_0'(n).$$

À partir de l'équation (20) et de la borne déduite on obtient

$$\mathbb{P}(\|\epsilon_A\theta - \epsilon_b\|_2 \geq \epsilon_2(\delta_n)) \leq \mathbb{P}(\|\epsilon_A\theta - \epsilon_b - \mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \geq \epsilon_2(\delta_n) - \epsilon_0'(n)) \leq \delta_n$$

pour $\epsilon_2(\delta_n)$ vérifiant—cf. lemme 2—

$$\epsilon_2(\delta_n) = \frac{2\sqrt{d}L\|\Delta_i\|_\infty}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{\left(\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil + 1\right)J(n-1,\delta_n)} + \frac{2\sqrt{d}L\|\Delta_i\|_\infty}{(n-1)(1-\lambda\gamma)}\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil + \epsilon_0'(n). \qquad (21)$$

Il nous reste à borner le terme $\|\Delta_i\|_\infty$. Pour cela, il suffit de borner $v_{LSTD(\lambda)}$. Pour tout $x \in \mathcal{X}$, on a

$$|v_{LSTD(\lambda)}(x)| = |\phi^T(x)\theta| \leq \|\phi^T(x)\|_2\|\theta\|_2 \leq \sqrt{d}L\|\theta\|_2,$$

où on obtient la première inégalité par l'inégalité de Cauchy-Schwarz. On peut borner $\|\theta\|_2$. En effet, on observe d'une part que

$$\|v_{LSTD(\lambda)}\|_\mu = \|\Phi\theta\|_\mu \geq \sqrt{\theta^T M_\mu\theta} \geq \sqrt{\nu}\|\theta\|_2.$$

D'autre part, on a

$$\|v_{LSTD(\lambda)}\|_\mu = \|(I - \Pi M)^{-1}\Pi(I - \lambda\gamma P)^{-1}r\|_\mu \leq \frac{R_{\max}}{1-\gamma} = V_{\max}.$$

Par conséquent

$$\|\theta\|_2 \leq \frac{V_{\max}}{\sqrt{\nu}}.$$

Il s'en suit que

$$\forall x \in \mathcal{X},\ |v_{LSTD(\lambda)}(x)| \leq \frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}}.$$

Ainsi, pour tout i on a

$$\begin{aligned}|\Delta_i| &= |v_{LSTD(\lambda)}(X_i) - \gamma v_{LSTD(\lambda)}(X_{i+1}) - r(X_i)| \\ &\leq \frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + \gamma\frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + (1-\gamma)V_{\max}.\end{aligned}$$

Puisque $\Phi^T D_\mu \Phi$ est une matrice symétrique, on a $\nu \leq \|\Phi^T D_\mu \Phi\|_2$. D'un autre côté on peut voir que

$$\|\Phi^T D_\mu \Phi\|_2 \leq d \max_{j,k} |\phi_k^t D_\mu \phi_j| = d \max_{j,k} |\phi_k^t D_\mu^{\frac{1}{2}} D_\mu^{\frac{1}{2}} \phi_j| \leq d \max_{j,k} \|\phi_k^t\|_\mu \|\phi_j\|_\mu \leq dL^2,$$

de sorte que $\nu \leq dL^2$. Il s'en suit, que pour tout i,

$$|\Delta_i| \leq \frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + \gamma\frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + \frac{\sqrt{d}L}{\sqrt{\nu}}(1-\gamma)V_{\max} = 2\frac{\sqrt{d}L}{\sqrt{\nu}}V_{\max}.$$

Conclusion de la preuve.

Nous allons maintenant conclure la preuve du théorème 1. Une fois qu'on a contrôlé les termes $\|\epsilon_A\|_2$ et $\|\epsilon_A\theta - \epsilon_b\|_2$, on peut déduire que

$$\begin{aligned}&\mathbb{P}\left\{\exists n \geq 1,\ \{\|\epsilon_A\|_2 \geq \epsilon_1(n,\delta_n)\} \cup \{\|\epsilon_A\theta - \epsilon_b)\|_2 \geq \epsilon_2(n,\delta_n)\}\right\} \\ \leq &\sum_{n=1}^{\infty} \mathbb{P}\left\{\|\epsilon_A\|_2 \geq \epsilon_1(n,\delta_n)\right\} + P\left\{\|\epsilon_A\theta - \epsilon_b)\|_2 \geq \epsilon_2(n,\delta_n)\right\} \\ \leq &2\sum_{n=1}^{\infty} \delta_n = \frac{1}{2}\frac{\pi^2}{6}\delta < \delta\end{aligned}$$

si on choisit $\delta_n = \frac{1}{4n^2}\delta$. D'après la seconde partie du lemme 4, pour tout δ, avec probabilité $1-\delta$, pour tout n tel que $\epsilon_1(n,\delta_n) < C$, on a $\hat{A}$ inversible et

$$\begin{aligned}\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_\mu &\leq \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}} \frac{\epsilon_2(n,\delta_n)}{1 - \frac{\epsilon_1(n,\delta_n)}{C}} \\ &= \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}}\left[\epsilon_2(n,\delta_n) + \frac{\epsilon_1(n,\delta_n)\,\epsilon_2(n,\delta_n)}{C - \epsilon_1(n,\delta_n)}\right].\end{aligned}$$

On obtiendra la borne dans le théorème 1 en remplaçant $\epsilon_1(n,\delta_n)$ et $\epsilon_2(n,\delta_n)$ par leurs définitions respectives dans les équations (19) et (21).

Pour compléter la preuve du théorème 1, il reste à montrer comment choisir C, ce qui nous permettra de montrer que la condition $\epsilon_1(n,\delta_n) < C < \frac{1}{\|A^{-1}\|_2}$ est équivalente à celle qui caractérise l'entier $n_0(\delta)$ définie dans le théorème 1. On a

$$\forall v \in \mathbb{R}^d, \|\Phi A^{-1} v\|_\mu = \sqrt{(A^{-1}v)^T M_\mu A^{-1} v} \geq \sqrt{\nu}\|A^{-1}v\|_2.$$

On peut observer que

$$\|\Phi A^{-1} v\|_\mu = \|(I - \Pi M)\Phi M_\mu^{-1} v\|_\mu \leq \frac{1-\lambda\gamma}{1-\gamma}\|\Phi M_\mu^{-1} v\|_\mu \leq \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}}\|v\|_2$$

où la dernière inégalité est obtenue à partir de l'équation (16). Par conséquent on a

$$\|A^{-1}\|_2 \leq \frac{1-\lambda\gamma}{(1-\gamma)\nu},$$

on peut prendre alors $C = \frac{(1-\gamma)\nu}{1-\lambda\gamma}$. Ceci conclut la preuve du théorème 1.

5 Conclusion et perspectives de travail

Cet article étudie la vitesse de convergence de l'algorithme LSTD(λ) en termes du nombre d'échantillons n et du paramètre λ. On a montré, sous l'hypothèse β-mélangeante, que la vitesse de convergence était de l'ordre $\tilde{O}(\frac{1}{\sqrt{n}})$. Pour cela, nous avons introduit une inégalité de concentration vectorielle pour les estimations basées sur des traces infiniment longues (lemme 2). Une version plus simple de cette inégalité de concentration, qui concerne de manière générale les processus exponentiellement β-mélangeants stationnaires (énoncée au lemme 6) pourrait être utile dans d'autres contextes où on aurait besoin de relâcher l'hypothèse iid sur les échantillons.

La borne de performance que nous déduisons de notre analyse est plus précise que celle obtenue par Lazaric *et al.* (2012) dans le cas $\lambda = 0$. L'analyse qu'ils ont employée utilise un modèle de régression Markovien. Nous pensons qu'il est possible, en utilisant la technique de troncature des traces que nous avons suivie ici, d'étendre leur analyse au cas $\lambda \neq 0$. Cependant, ce faisant, on gardera toujours le facteur $4\sqrt{2}$ dans la borne finale. On envisage, dans l'avenir, d'instancier la nouvelle borne obtenue dans le contexte itérations sur les politiques, comme il a déjà été fait pour $\lambda = 0$ dans (Lazaric *et al.*, 2012).

Une perspective intéréssante serait également d'étendre cette analyse dans le cas des politiques non stationnaires, Scherrer & Lesner (2012) ayant montré qu'elles amélioraient la borne de performance de l'algorithme itérations sur les politiques. Finalement une question un peu plus difficile serait d'envisager une telle analyse pour l'algorithme off-policy LSTD(λ), dont la convergence vient d'être récemment prouvée par Yu (2010).

A Preuve du Lemme 2

Soient

$$\epsilon_1 = \frac{1}{n-1}\sum_{i=1}^{m-1} G_i - \mathbb{E}[G_i]$$

$$\text{et} \quad \epsilon_2 = \frac{1}{n-1}\sum_{i=m}^{n-1}(z_i - z_i^m)\tau(X_i, X_{i+1})^T - \mathbb{E}[(z_i - z_i^m)\tau(X_i, X_{i+1})^T].$$

On a

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n-1} G_i - \mathbb{E}[G_i] &= \frac{1}{n-1}\sum_{i=m}^{n-1} G_i - \mathbb{E}[G_i] + \epsilon_1 \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1} z_i(\tau(X_i, X_{i+1})^T - \mathbb{E}[z_i(\tau(X_i, X_{i+1})^T] + \epsilon_1 \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1} z_i^m\tau(X_i, X_{i+1})^T - \mathbb{E}[z_i^m\tau(X_i, X_{i+1})^T] + \epsilon_1 + \epsilon_2 \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m]) + \epsilon_1 + \epsilon_2. \qquad (22)
\end{aligned}$$

Pour tout i, on a $\|z_i\|_\infty \leq \frac{L}{1-\lambda\gamma}$, $\|G_i\|_\infty \leq \frac{LL'}{1-\lambda\gamma}$, et $\|z_i - z_i^m\|_\infty \leq \frac{(\lambda\gamma)^m L}{1-\lambda\gamma}$. Par conséquent—en utilisant $\|M\|_2 \leq \|M\|_F = \sqrt{d \times k}\|x\|_\infty$ pour $M \in \mathbb{R}^{d\times k}$ avec x le vecteur obtenu en concaténant les colonnes de M—, on obtient

$$\|\epsilon_1 + \epsilon_2\|_2 \leq \frac{2(m-1)\sqrt{d \times k}LL'}{(n-1)(1-\lambda\gamma)} + \frac{2(\lambda\gamma)^m\sqrt{d \times k}LL'}{(1-\lambda\gamma)}. \qquad (23)$$

En concaténant les colonnes de la matrice G_i^m, la matrice peut être considérée comme un vecteur U_i^m de taille dk. On a alors pour tout $\epsilon > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m])\right\|_2 \geq \epsilon\right) \leq \mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m])\right\|_F \geq \epsilon\right)$$
$$= \mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(U_i^m - \mathbb{E}[U_i^m])\right\|_2 \geq \epsilon\right). \quad (24)$$

Les variables U_i^m définissant un processus β-mélangeant stationnaire (lemme 1), on utilise la technique de décomposition proposée par Yu (1994) qui consiste à regrouper les variables $U_m^m, \ldots, U_{n-1}^m$ en $2\mu_{n-m}$ blocs de taille a_{n-m} (on suppose $n - m = 2a_{n-m}\mu_{n-m}$). Les blocs sont de deux sortes : ceux qui contiennent les indices pairs $E = \cup_{l=1}^{\mu_{n-m}} E_l$ et ceux qui contiennent les indices impairs $H = \cup_{l=1}^{\mu_{n-m}} H_l$. En regroupant les variables dans des blocs on obtient

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq \mathbb{P}\left(\left\|\sum_{i\in H}U_i^m - \mathbb{E}[U_i^m]\right\|_2 + \left\|\sum_{i\in E}U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq (n-m)\frac{\epsilon}{2}\right) \quad (25)$$
$$\leq \mathbb{P}\left(\left\|\sum_{i\in H}U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) + \mathbb{P}\left(\left\|\sum_{i\in E}U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) \quad (26)$$
$$= 2\mathbb{P}\left(\left\|\sum_{i\in H}U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right). \quad (27)$$

L'équation (25) découle de l'inégalité triangulaire. L'équation (26) découle du fait que $\{X + Y \geq a\}$ implique $\{X \geq \frac{a}{2}\}$ ou $\{Y \geq \frac{a}{2}\}$. La stationnarité du processus implique l'équation (27). Puisque $H = \cup_{l=1}^{\mu_{n-m}} H_l$ on a

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}}\sum_{i\in H_l}U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right)$$
$$= 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}}U(H_l) - \mathbb{E}[U(H_l)]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) \quad (28)$$

où on définit $U(H_l) = \sum_{i\in H_l} U_i^m$. Considérons maintenant la séquence de blocs $(U'(H_l))_{l=1,\ldots,\mu_{n-m}}$ indépendants tel que chaque bloc $U'(H_l)$ admet la même distribution que $U(H_l)$. On va utiliser le lemme suivant.

Lemme 5. *Yu (1994) Soit $X_1, \ldots, X_n$ une séquence d'échantillons générés à partir d'un processus β-mélangeant de coefficient $\{\beta_i\}$. Soit $X(H) = (X(H_1), \ldots, X(H_{\mu_{n-m}}))$ tel que pour tout j, $X(H_j) = (X_i)_{i\in H_j}$. On définit $X'(H) = (X'(H_1), \ldots, X'(H_{\mu_{n-m}}))$ tel que les variables $X'(H_j)$ sont indépendantes et tel que pour tout j, $X'(H_j)$ a la même distribution que $X(H_j)$. Soient Q et Q' les distributions de $X(H)$ et $X'(H)$ respectivement. Pour toute fonction mesurable $h : \mathcal{X}^{a_n\mu_n} \to \mathbb{R}$ bornée par B, on a*

$$|\mathbb{E}_Q[h(X(H))] - \mathbb{E}_{Q'}[h(X'(H))]| \leq B\mu_n\beta_{a_n}.$$

En appliquant le Lemme 5, l'équation (28) implique que :

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}}U'(H_l) - \mathbb{E}[U'(H_l)]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) + 2\mu_{n-m}\beta_{a_{n-m}}. \quad (29)$$

Les variables $U'(H_l)$ sont indépendantes, de plus le processus $\sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)]$ est une $\sigma(U'(H_1), \dots, U'(H_{\mu_{n-m}}))$ martingale :

$$\begin{aligned}
&\mathbb{E}\left[\sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)] \,\middle|\, U'(H_1), \dots, U'(H_{\mu_{n-m}-1})\right] \\
&= \sum_{l=1}^{\mu_{n-m}-1} U'(H_l) - \mathbb{E}[U'(H_l)] + \mathbb{E}[U'_{H_{\mu_{n-m}}} - \mathbb{E}[U'_{H_{\mu_{n-m}}}]] \\
&= \sum_{l=1}^{\mu_{n-m}-1} U'(H_l) - \mathbb{E}[U'(H_l)].
\end{aligned}$$

On peut alors appliquer l'inégalité de concentration suivante.

Lemme 6 (Hayes (2005))**.** *Soit $X = (X_0, \dots, X_n)$ une martingale discrète prenant ses valeurs dans un espace euclidien telle que $X_0 = 0$ et pour tout i, $\|X_i - X_{i-1}\|_2 \leq B_2$ presque sûrement. Pour tout ϵ,*

$$P\left\{\|X_n\|_2 \geq \epsilon\right\} < 2e^2 e^{-\frac{\epsilon^2}{2n(B_2)^2}}.$$

En effet en prenant $X_{\mu_{n-m}} = \sum_{l=}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)]$, et en observant que $\|X_i - X_{i-1}\| = \|U'(H_l) - \mathbb{E}[U'(H_l)]\|_2 \leq a_{n-m}C$ avec $C = \frac{2\sqrt{dk}LL'}{1-\lambda\gamma}$, le Lemme 6 implique

$$\begin{aligned}
\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) &\leq 2e^2 e^{-\frac{(n-m)^2\epsilon^2}{32\mu_{n-m}(a_{n-m}C)^2}} \\
&= 2e^2 e^{-\frac{(n-m)\epsilon^2}{16a_{n-m}C^2}}
\end{aligned}$$

où la deuxième inégalité est obtenue en utilisant l'égalité $2a_{n-m}\mu_{n-m} = n - m$. En combinant les équations (28) et (29), on obtient finalement

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 4e^2 e^{-\frac{(n-m)\epsilon^2}{16a_{n-m}C^2}} + 2(n-m)\beta^U_{a_{n-m}}.$$

Le vecteur U_i^m est une fonction du vecteur $Z_i = (X_{i-m+1}, \dots, X_{i+1})$. D'après le lemme 1 on sait que pour tout $j > m$,

$$\beta_j^U \leq \beta_j^Z \leq \beta_{j-m}^X \leq \overline{\beta}e^{-b(j-m)^\kappa}.$$

On obtient alors

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 4e^2 e^{-\frac{(n-m)\epsilon^2}{16a_{n-m}C^2}} + 2(n-m)\overline{\beta}e^{-b(a_{n-m}-m)^\kappa} = \delta'. \quad (30)$$

Pour avoir le même exposant dans les deux exponentielles, on va suivre le même raisonnement que Lazaric *et al.* (2012) ; En prenant $a_{n-m} - m = \left\lceil \frac{C_2(n-m)\epsilon^2}{b} \right\rceil^{\frac{1}{\kappa+1}}$ avec $C_2 = (16C^2\zeta)^{-1}$, et $\zeta = \frac{a_{n-m}}{a_{n-m}-m}$, on obtient

$$\delta' \leq (4e^2 + (n-m)\overline{\beta})\exp\left(-\min\left\{\left(\frac{b}{(n-m)\epsilon^2 C_2}\right), 1\right\}^{\frac{1}{k+1}} \frac{1}{2}(n-m)C_2\epsilon^2\right). \quad (31)$$

En écrivant

$$\Lambda(n,\delta) = \log\left(\frac{2}{\delta}\right) + \log(\max\{4e^2, n\overline{\beta}\})$$

et

$$\epsilon(\delta) = \sqrt{2\frac{\Lambda(n-m,\delta)}{C_2(n-m)}\max\left\{\frac{\Lambda(n-m,\delta)}{b}, 1\right\}^{\frac{1}{\kappa}}},$$

on peut montrer que

$$\exp\left(-\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2C_2}\right),1\right\}^{\frac{1}{k+1}}\frac{1}{2}(n-m)C_2(\epsilon(\delta))^2\right)\leq\exp\left(-\Lambda(n-m,\delta)\right). \quad (32)$$

En effet [11], il y a deux cas :

1. Supposons $\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2C_2}\right),1\right\}=1$. On a

$$\begin{aligned}&\exp\left(-\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2C_2}\right),1\right\}^{\frac{1}{k+1}}\frac{1}{2}(n-m)C_2(\epsilon(\delta))^2\right)\\ =&\exp\left(-\Lambda(n-m,\delta)\max\left\{\frac{\Lambda(n-m,\delta)}{b},1\right\}^{\frac{1}{k}}\right)\\ \leq&\exp\left(-\Lambda(n-m,\delta)\right).\end{aligned}$$

2. Supposons maintenant que $\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2C_2}\right),1\right\}=\left(\frac{b}{(n-m)(\epsilon(\delta))^2C_2}\right)$. On a alors

$$\begin{aligned}\exp\left(-\frac{1}{2}b^{\frac{1}{k+1}}((n-m)C_2(\epsilon(\delta))^2)^{\frac{k}{k+1}}\right)&=\exp\left(-\frac{1}{2}b^{\frac{1}{k+1}}(\Lambda(n-m,\delta)^{\frac{k}{k+1}}\max\left\{\frac{\Lambda(n-m,\delta)}{b},1\right\}^{\frac{1}{k+1}}\right)\\ &=\exp\left(-\frac{1}{2}\Lambda(n-m,\delta)^{\frac{k}{k+1}}\max\left\{\Lambda(n-m,\delta),b\right\}^{\frac{1}{k+1}}\right)\\ &\leq\exp\left(-\Lambda(n-m,\delta)\right).\end{aligned}$$

En combinant les équations (31) et (32), on trouve

$$\delta'\leq(4e^2+(n-m)\overline{\beta})\exp\left(-\Lambda(n-m,\delta)\right).$$

En remplaçant $\Lambda(n-m,\delta)$ par son expression, on obtient

$$\exp\left(-\Lambda(n-m,\delta)\right)=\frac{\delta}{2}\max\{4e^2,(n-m)\overline{\beta}\}^{-1}.$$

Puisque $4e^2\max\{4e^2,(n-m)\overline{\beta}\}^{-1}\leq 1$ et $(n-m)\overline{\beta}\max\{4e^2,(n-m)\overline{\beta}\}^{-1}\leq 1$, on a

$$\delta'\leq 2\frac{\delta}{2}\leq\delta.$$

Puisque $a_{n-m}-m\geq 1$, on a

$$\zeta=\frac{a_{n-m}}{a_{n-m}-m}=\frac{a_{n-m}-m+m}{a_{n-m}-m}\leq 1+m.$$

Soit $J(n,\delta)=32\Lambda(n,\delta)\max\left\{\frac{\Lambda(n,\delta)}{b},1\right\}^{\frac{1}{\kappa}}$. L'équation (30) est réduite à

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(U_i^m-\mathbb{E}[U_i^m])\right\|_2\geq\frac{C}{\sqrt{n-m}}\left((m+1)J(n-m,\delta)\right)^{\frac{1}{2}}\right)\leq\delta. \quad (33)$$

$J(n,\delta)$ est une fonction croissante de n, et $\frac{n-1}{\sqrt{n-1}(n-m)}=\frac{1}{\sqrt{n-m}}\sqrt{\frac{n-1}{n-m}}\geq\frac{1}{\sqrt{n-m}}$, on a alors

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{n-1}\sum_{i=m}^{n-1}(G_i^m-\mathbb{E}[G_i^m])\right\|_2\geq\frac{C}{\sqrt{n-1}}\left((m+1)J(n-1,\delta)\right)^{\frac{1}{2}}\right)\\ \leq&\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m-\mathbb{E}[G_i^m])\right\|_2\geq\frac{C}{\sqrt{n-1}}\frac{n-1}{n-m}\left((m+1)J(n-1,\delta)\right)^{\frac{1}{2}}\right)\\ \leq&\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m-\mathbb{E}[G_i^m])\right\|_2\geq\frac{C}{\sqrt{n-m}}\left((m+1)J(n-m,\delta)\right)^{\frac{1}{2}}\right).\end{aligned}$$

11. Cette inégalité existe dans Lazaric *et al.* (2012), et est développée ici pour être complet.

En utilisant les équations (24) et (33), on peut déduire que

$$\mathbb{P}\left(\left\|\frac{1}{n-1}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m])\right\|_2 \geq \frac{C}{\sqrt{n-1}}\left((m+1)J(n-1,\delta)\right)^{\frac{1}{2}}\right) \leq \delta. \tag{34}$$

En combinant les équations (22), (23), (34), en utilisant $C = \frac{2\sqrt{dk}LL'}{1-\lambda\gamma}$, et en choisissant $m = \left\lceil \frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)} \right\rceil$, on obtient le résultat.

Références

ARCHIBALD T., MCKINNON K. & THOMAS L. (1995). On the generation of Markov decision processes. *Journal of the Operational Research Society*, **46**, 354–361.

BERTSEKAS D. & TSITSIKLIS J. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

BOYAN J. A. (2002). Technical update : Least-squares temporal difference learning. *Machine Learning*, **49**(2–3), 233–246.

HAYES T. P. (2005). A large-deviation inequality for vector-valued martingales. Manuscript.

LAZARIC A., GHAVAMZADEH M. & MUNOS R. (2012). Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, **13**, 3041–3074.

NEDIC A. & BERTSEKAS D. P. (2002). Least squares policy evaluation algorithms with linear function approximation. *Theory and Applications*, **13**, 79–110.

SCHERRER B. (2010). Should one compute the temporal difference fix point or minimize the bellman residual ? the unified oblique projection view. In *ICML*.

SCHERRER B. & LESNER B. (2012). On the use of non-stationary policies for stationary infinite-horizon Markov decision processes. In *NIPS 2012 Adv.in Neural Information Processing Systems*, South Lake Tahoe, United States.

SZEPESVÁRI C. (2010). *Algorithms for Reinforcement Learning*. Morgan and Claypool.

TSITSIKLIS J. N. & ROY B. V. (1997). *An analysis of temporal-difference learning with function approximation*. Rapport interne, IEEE Transactions on Automatic Control.

YU B. (1994). Rates of convergence for empirical processes stationnary mixing consequences. *The Annals of Probability*, **19**, 3041–3074.

YU H. (2010). Convergence of least-squares temporal difference methods under general conditions. In *ICML*.